Structured Credit Assignment in Mice

Kevin J. Miller (kevinjmiller@deepmind.com)
DeepMind and University College London
London, United Kingdom

Laurence Freeman (laurence.freeman.19@ucl.ac.uk)
University College London
London, United Kingdom

Yu Jin Oh (yu.oh.20@alumni.ucl.ac.uk)
University College London
London, United Kingdom

Matthew Botvinick (botvinick@deepmind.com)
DeepMind
London, United Kingdom

Kenneth Harris (kenneth.harris@ucl.ac.uk)
University College London
London, United Kingdom

Abstract

Reinforcement learning requires associating rewards with one or more of the states or actions that preceded them. The question of exactly which states or actions to associate with each reward is referred to as the "credit assignment problem", and better solutions result in more efficient learning. In humans, credit assignment is informed by knowledge of the causal structure of the world. Here, we adapt a "structured" credit assignment task from the human literature for use with head-fixed mice. In this task, rewards of one type ("controllable") depend causally on the mouse's actions, while another distinguishable type ("distractor") is independent of those actions. We present behavioral evidence that mice, like humans, adopt a strategy that is partially structure-sensitive: they update their behavior based on rewards of both types, but they update more strongly to the controllable reward. This work opens the door to investigations of the neural mechanisms of structured credit assignment using the wide range of tools that are available in head-fixed mice.

Introduction

Credit assignment in artificial agents typically uses a recency heuristic: credit for a reward is assigned to all states and actions that preceded it recently. This heuristic has intuitive appeal, however it is limiting in certain types of situations: for example when a long time delay occurs between an action and the reward that it causes, or when there are structured relationships governing which types of actions are able to cause which types of rewards. Developing new credit assignment methods for these and other situations is an active area of machine learning research (Harutyunyan et al., 2019; Mesnard et al., 2020; Chelu et al., 2022).

The brain's strategies for credit assignment remain largely unknown. While there are data indicating that recency indeed plays a role (Yagishita et al., 2014; Lehmann et al., 2019), there is also evidence that causal task structure is taken into account (Gershman, Pesaran, and Daw, 2009; Jocham et al., 2016; Moran, Dayan, and Dolan, 2021). This "structured" credit assignment is typically studied in humans, limiting the range of neuroscience tools that can be applied. Here, we adapt a structured credit assignment task from the human literature, the "distractor rewards" task (Jocham et al., 2016), for use with head-fixed mice. We find that mice, like humans, learn using both recency-based as well as structured credit assignment.

Results

Distractor rewards task for mice

Mice perform the task while head-fixed with their forepaws on a wheel, indicating their choice on each trial by rotating the wheel either to the left or to the right (Burgess et al., 2017). The rig contains two reward spouts, each containing sweet liquid reward of a different flavor (cherry and grape Kool-Aid; Figure 1a). After the mouse makes its choice, each of these spouts delivers its reward with a certain probability. Reward probabilities are independent between the two spouts, so in total there are four possible outcomes: no reward, cherry only, grape only, or both (Figure 1b). For each mouse, one flavor is designated the "controllable" flavor, while the other is designated the "distractor". The contingencies of the controllable flavor follow a probabilistic reversal learning schedule: on each trial one action is rewarded with high probability while the other is rewarded with low probability, and these probabilities swap at unpredictable intervals. With respect to the distractor flavor, the reward probability is constant on all trials, regardless of the mouse's choice (Figure 1c).

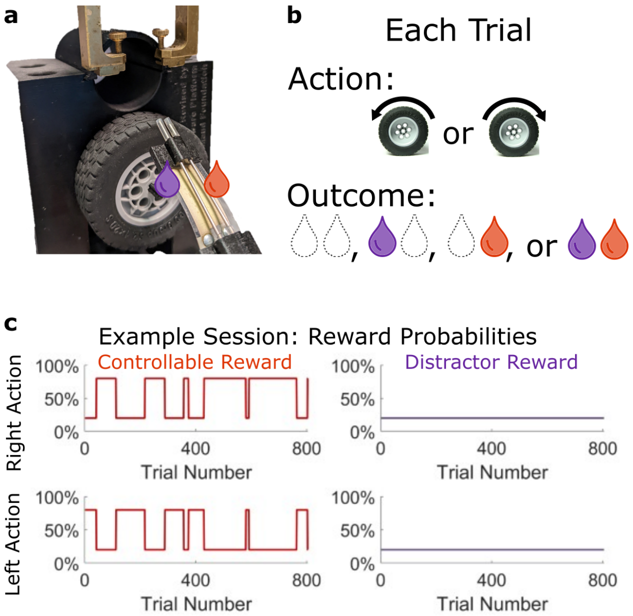

Figure One: Distractor rewards task for mice

This difference between controllable and distractor rewards allows our task to assay structure in credit assignment. We expect an agent that does not use task structure to show noncontingent recency-based credit assignment: reinforcing actions that are followed by a reward and switching away from actions that are not, regardless of that reward's flavor. In contrast, we expect an agent that uses structured credit assignment to reinforce only actions that are followed by reward a controllable reward, but switch away from those that were not, regardless of the presence or absence of the distractor reward.

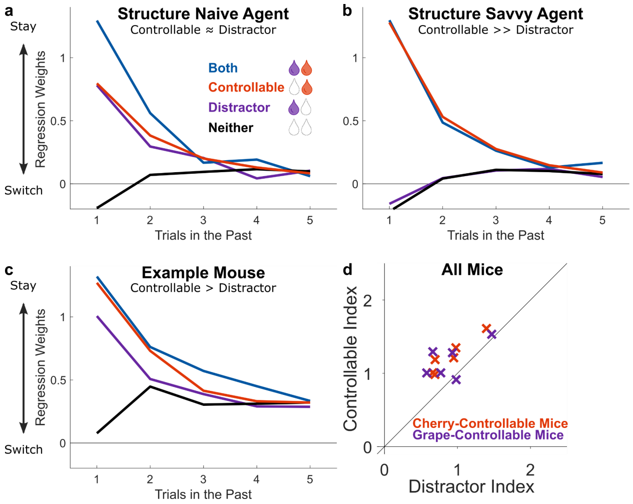

Figure Two: Mice show structured as well as noncontingent credit assignment

We characterize these behavioral patterns using a trial-history regression model (Lau & Glimcher, 2005). This model fits separate weights for each outcome type (none, controllable, distractor, both), which quantify the tendency to repeat (vs. switch away from) recent actions that were followed by that outcome. We first apply this approach to two synthetic datasets: one from a "structure naive" agent that learns based on recency alone (Figure 2a), the other from a "structure savvy" agent that learns only from the controllable reward (Figure 2b). The fit regression weights show the expected patterns: the structure-naive agent earns large positive weights for all three rewarding outcomes and negative weights for the no-reward outcome, while the structure-savvy agent shows positive weights for the "controllable" and "both" outcomes and negative weights for the "distractor" and "no-reward" outcomes. Next, we apply this analysis to a mouse behavioral dataset (198 behavioral sessions from twelve mice), fitting a separate set of weights for each mouse (Figure 2c). We summarize the data for each mouse by computing a "controllable index" (the total difference between the weights for the controllable outcome and the no-reward outcome) as well as a "distractor index" (the total difference between the weights for the distractor outcome and the no-reward outcome). For mice, the controllable index was larger than the distractor index (Figure 2d, points are above the diagonal), indicating that they took task structure into account when performing credit assignment. At the same time, the distractor index was greater than zero (Figure 2d, points are to the right of the vertical axis), indicating that mice also perform noncontingent credit assignment. We checked both of these results using a conditional randomization test that allows inferring causal effects despite the statistical issues arising from correlated timeseries (Harris, 2020), and found them to be statistically significant (p<0.001).

Conclusions and Directions

We found that mice, like humans, use both structured as well as noncontingent credit assignment in the distractor rewards task. This opens the door to investigations of the neural mechanisms of credit assignment using the wide range of tools that are available for head-fixed mice. We are in the process of collecting a dataset of electrophysiological recordings using high-density Neuropixels probes (Jun et al., 2017) surveying both frontal cortex (secondary motor, anterior cingulate, prelimbic, infralimbic, medial orbital, ventral orbital, and lateral orbital) and striatum (caudoputamen and nucleus accumbens), the details of which will be reported elsewhere.

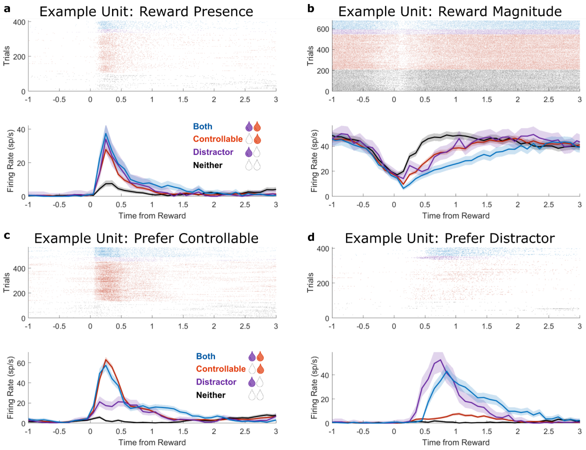

Figure Three: Example reward responses from units in frontal cortex

Acknowledgements

This work was supported by funding from the Wellcome Trust (205093/Z/16/Z) and from the European Research Council (694401).

References

Burgess, C. P., Lak, A., Steinmetz, N. A., Zatka-Haas, P., Bai Reddy, C., Jacobs, E. A. K., Linden, J. F., Paton, J. J., Ranson, A., Schröder, S., Soares, S., Wells, M. J., Wool, L. E., Harris, K. D., & Carandini, M. (2017). High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice. *Cell Reports*, *20*(10), 2513–2524.

Chelu, V., Borsa, D., Precup, D., & van Hasselt, H. (2022). Selective credit assignment. In *arXiv [cs.LG]*. arXiv.

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(43), 13524–13531.

Harris, K. D. (2020). Nonsense correlations in neuroscience. *bioRxiv*

Harutyunyan, A., Dabney, W., Mesnard, T., Gheshlaghi Azar, M., Piot, B., Heess, N., van Hasselt, H. P., Wayne, G., Singh, S., Precup, D., & Others. (2019). Hindsight credit assignment. *Advances in Neural Information Processing Systems*, *32*.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., Rushworth, M. F. S., & Behrens, T. E. J. (2016). Reward-Guided Learning with and without Causal Attribution. *Neuron*, *90*(1), 177–190.

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç., Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., … Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232–236.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579.

Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019). One-shot learning and behavioral eligibility traces in sequential decision making. *eLife*, *8*.

Mesnard, T., Weber, T., Viola, F., Thakoor, S., Saade, A., Harutyunyan, A., Dabney, W., Stepleton, T., Heess, N., Guez, A., Moulines, É., Hutter, M., Buesing, L., & Munos, R. (2020). Counterfactual Credit Assignment in Model-Free Reinforcement Learning.

Moran, R., Dayan, P., & Dolan, R. J. (2021). Human subjects exploit a cognitive map for credit assignment. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(4).

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(6204), 1616–1620.